Journal of Biotechnology 158 (2012) 151–155

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Short communication

Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines

Matthias Hackl[a,1], Vaibhav Jadhav[a,1], Tobias Jakobi[b], Oliver Rupp[b], Karina Brinkrolf[b], Alexander Goesmann[b], Alfred Pühler[b], Thomas Noll[c], Nicole Borth[a,d], Johannes Grillari[a,*]

[a] Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria
[b] Centrum für Biotechnologie, Universität Bielefeld, 33594 Bielefeld, Germany
[c] AG Zellkulturtechnik, Technische Fakultät, Universität Bielefeld, 33549 Bielefeld, Germany
[d] ACIB GmbH, Austrian Centre of Industrial Biotechnology, Graz, Austria

ARTICLE INFO

Article history:
Received 15 November 2011
Received in revised form 11 January 2012
Accepted 13 January 2012
Available online 25 January 2012

Keywords:
MicroRNA
microRNA stemloops
Chinese hamster ovary
Cell engineering

ABSTRACT

MicroRNAs (miRNAs) have recently entered Chinese hamster ovary (CHO) cell culture technology, due to their severe impact on the regulation of cellular phenotypes. Applications of miRNAs that are envisioned range from biomarkers of favorable phenotypes to cell engineering targets. These applications, however, require a profound knowledge of miRNA sequences and their genomic organization, which exceeds the currently available information of ~400 conserved mature CHO miRNA sequences. Based on these recently published sequences and two independent CHO-K1 genome assemblies, this publication describes the computational identification of CHO miRNA genomic loci. Using BLAST alignment, 415 previously reported CHO miRNAs were mapped to the reference genomes, and subsequently assigned to a distinct genomic miRNA locus. Sequences of the respective precursor-miRNAs were extracted from both reference genomes, folded *in silico* to verify correct structures and cross-compared. In the end, 212 genomic loci and pre-miRNA sequences representing 319 expressed mature miRNAs (approximately 50% of miRNAs represented matching pairs of 5′ and 3′ miRNAs) were submitted to the miRBase miRNA repository. As a proof-of-principle for the usability of the published genomic loci, four likely polycistronic miRNA cluster were chosen for PCR amplification using CHO-K1 and DHFR (-) genomic DNA. Overall, these data on the genomic context of miRNA expression in CHO will simplify the development of tools employing stable overexpression or deletion of miRNAs, allow the identification of miRNA promoters and improve detection methods such as microarrays.

Chinese hamster ovary (CHO) cells are currently the first choice mammalian cell line for the production of complex therapeutic proteins requiring proper folding and post-translational modifications, creating an annual revenue exceeding 100 billion USD (Mudhar, 2006). With the publication of a CHO-K1 draft genome (Xu et al., 2011), as well as thorough analysis of the CHO mRNA transcriptome (Becker et al., 2011), the basis for genomic characterization of CHO cells has been set and will allow the development of novel tools to rationally design CHO cells as bioindustrial work horses.

Therefore, microRNAs (miRNAs) have been discussed as promising tools for CHO cell characterization as well as engineering (Barron et al., 2011). This family of small non-coding RNAs, which by now encompasses more than 1000 sequences for mouse and human (Griffiths-Jones, 2010), acts by negative regulation of gene expression due to post-transcriptional repression of mRNA translation (Hüttenhofer and Schattner, 2006). The ~22 nt long mature miRNAs that catalyze this repression are the result of enzymatic processing of a primary RNA-Polymerase II miRNA transcript: in the nucleus, RNase III Drosha together with Dgcr8 cleave a ~70 nt long single-stranded RNA referred to as precursor miRNA (pre-miR) or miRNA hairpin/stem-loop due to its characteristic secondary structure (Gregory et al., 2004). Pre-miRNAs are exported into the cytoplasm where cleavage of the loop by the RNase Dicer generates a duplex of two ~22 nt long mature miRNAs (Takeshita et al., 2007). The partial sequence complementarity underlying the miRNA:mRNA interaction, allows single miRNAs to bind up to 100 distinct mRNAs (Selbach et al., 2008), thus potentially orchestrating the expression of whole gene networks similar to transcription factors. This range in target regulation achieved by individual miRNAs is mirrored in their biological relevance, which includes control of cellular proliferation and energy metabolism as well as stress resistance and cell death (Müller et al., 2008).

* Corresponding author. Tel.: +43 1 47654 6230.
E-mail address: johannes.grillari@boku.ac.at (J. Grillari).
[1] These authors contributed equally.

0168-1656 © 2012 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.jbiotec.2012.01.019

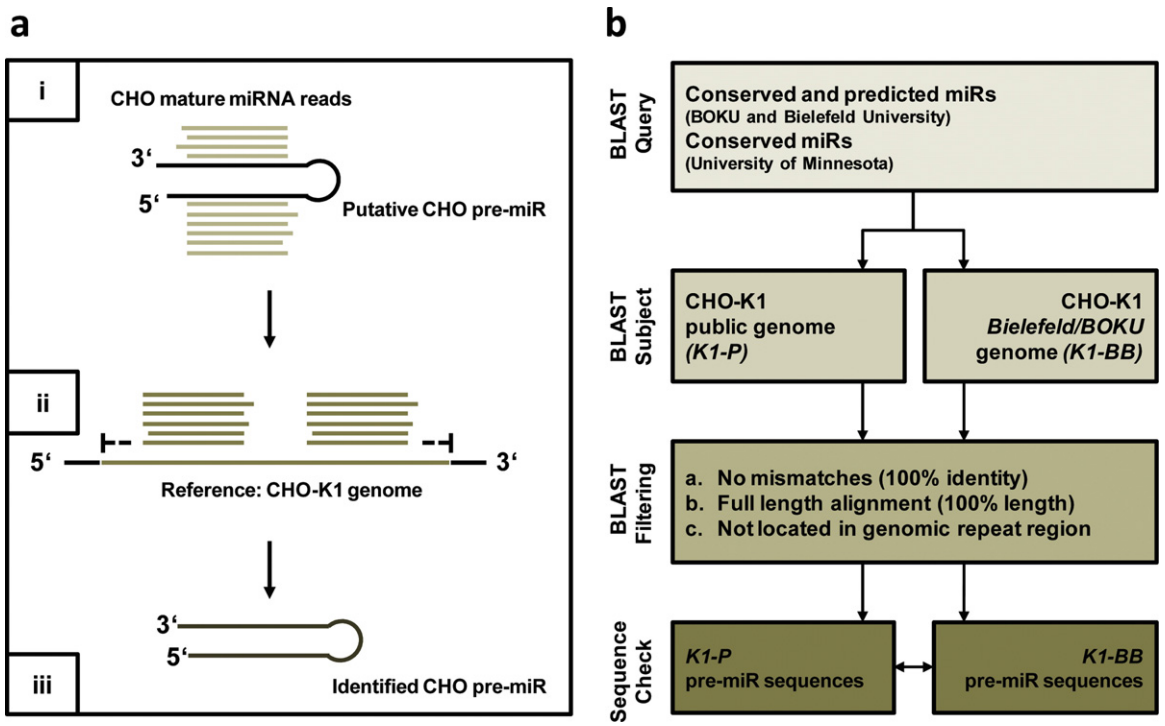

Fig. 1. Strategy for identification of CHO pre-miR sequences from genomic references. (a) Schematic outline of identification strategy. (b) Flow chart illustrating the sequence identification strategy in detail: all currently published CHO mature miRNA sequences were BLAST-aligned to two independent CHO-K1 genomic reference sequence assemblies (*K1-P and K1-BB*). BLAST results were filtered for alignments with zero mismatches (100% identity) and alignment lengths equal to the mature miRNA length (100% length). Additionally, miRNAs mapping to genomic repeat regions were removed. From the remaining genomic loci, the respective pre-miRNA sequences were extracted independently from both genomic references and cross-checked.

Two studies have so far addressed the identification and annotation of CHO miRNAs, and independently reported the expression of 350 (Johnson et al., 2011) and 365 (Hackl et al., 2011) mature miRNAs, but have not identified the respective genomic loci or pre-miRNA sequences. This information is, however, necessary for (i) mimicking endogenous miRNA expression, since pre-miRNA secondary structures can be target to regulation of miRNA stability (Michlewski et al., 2008), for (ii) understanding transcriptional regulation of specific miRNAs, as well as for (iii) phylogenetic analyses.

Based on the alignment of a combined set of previously reported mature miRNA sequences against two independent CHO-K1 genome reference sequences, we here report the identification of miRNA gene loci and extraction of the respective pre-miRNA sequences from both genomic references, followed by cross-comparison of the derived sequences (Fig. 1). In detail, the employed strategy used two public datasets containing sequences of mature CHO miRNAs with expression levels detectable by next-generation sequencing (Johnson et al., 2011; Hackl et al., 2011). Both datasets were downloaded, reduced by redundant isomiR sequences as well as recently reported non-coding RNAs in miRBase version 18.0 (Griffiths-Jones, 2010), and then merged into one dataset containing 415 miRNAs of which 22 were putative novel miRNA sequences. These sequences were further used as "query" (given in Supplementary Data 1) for BLAST alignment against two distinct CHO genome references using blastn with nucleotide mismatch penalty −2, and nucleotide match reward +1. The first reference consisted of the recently published CHO-K1 sequence (Xu et al., 2011) hereafter referred to as "*K1-P*" (for "public") and the second reference being a low coverage, so far unpublished CHO-K1 genome assembly by Bielefeld University and BOKU University referred to as "*K1-BB*" (Table 1). In brief, the K1-BB genome (ATCC CCL-61) was sequenced on an Illumina Genome Analyser IIx in a 2 × 125 bp sequencing run on six lanes according to the manufacturer's manuals. Sequencing resulted in 411 million reads and 51 Gbp which leads to an estimated genome coverage of 17-fold considering a genome size of 3 Gbp. Assembly of the sequence data was performed with velvet 1.0.4 resulting in 11.4 million contigs that can be downloaded at ftp://ftp.cebitec.uni-bielefeld.de/pub/supplements/2011/Hackl_JBiotech/.

Following filtering of BLAST alignments (Fig. 1), a total of 365 out of 415 distinct mature miRNAs could be mapped to either genomic reference. In detail, 353 distinct mature miRNAs gave a perfect BLAST hit against the *K1-P* reference, while 330 miRNAs could be aligned to the *K1-BB* reference with an overlap of 318 miRNAs, shown as Venn diagram in Fig. 2a (Hulsen et al., 2008). While the majority of miRNAs exhibited a single exact match in the reference genome, some miRNAs exhibited two or more exact matches (Fig. 2b). This might have biological reasons, since duplications of miRNA genes are known to result in 100% identical paralogous sequences present in other parts of the genome (Gardner et al., 2009). Alternatively, the observed multiple hits could be a consequence of incomplete assembly of the genomic references. This would explain the reduction in multiple perfect matches from the incompletely assembled 2.9 Gbp *K1-BB* genome to the almost completely assembled 2.45 Gbp *K1-P* genome from 28% to 16% of the aligned miRNAs (Fig. 2b). Nevertheless, 15 miRNAs exhibited more

Table 1
Genome references for identification of CHO pre-miR sequences.

	K1-P	K1-BB
Genome size (Gbp)	2.40	2.98
Contigs	109,151	11,400,490
Average contig length	21,986	261
Median contig length	503	124.5
x Coverage	95	17.1

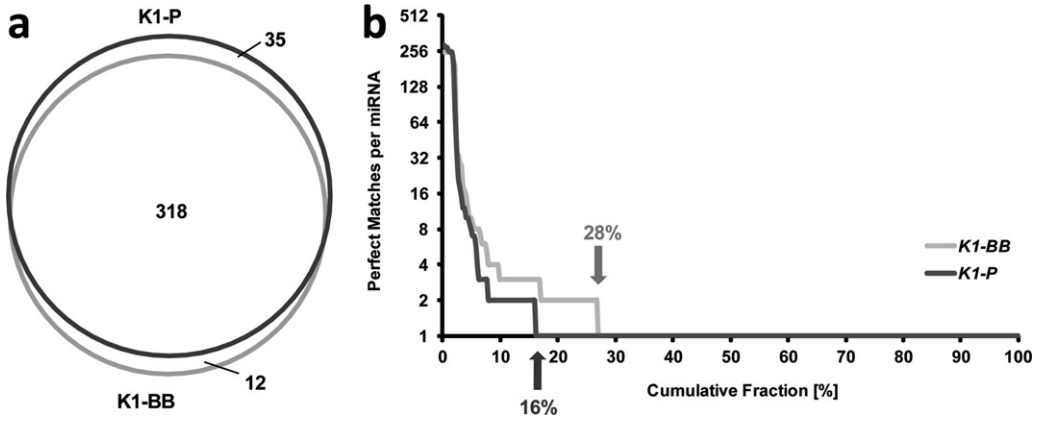

Fig. 2. BLAST alignment of mature miRNAs to two different genomic reference sequences. (a) Size-adjusted Venn diagram indicating that 318 mature miRNAs were aligned to both reference genomes, while 35 and 12 mature miRNAs could only be aligned to *K1-P* or *K1-BB*, respectively. (b) The cumulative fraction of BLAST-aligned miRNAs is plotted against the number of perfect genomic matches identified; for each miRNA; 16% and 28% of miRNAs could be perfectly aligned to two or more genomic locations in the *K1-P* (black) or *K1-BB* (gray) genomic reference sequence.

than 10 and up to 250 perfect matches (Supplementary Table 1), which indicates that these are repeat derived small RNAs rather than canonical miRNAs. Hence, these miRNAs were removed from the BLAST results and not considered for further analysis as well as submission to miRBase.

In the next step, the genomic locations of BLAST aligned mature miRNAs were analyzed in detail to identify the respective pre-miRNA sequences (Fig. 3a): genomic locations where two miRNAs could be aligned in close proximity indicate miRNA genes from which two mature miRNAs – corresponding to the 5′ and 3′ miRNA – are produced. Other genomic loci were mapped by only one miRNA, suggesting the expression of just one active mature miRNA, which is either derived from the 5′ or 3′ arm of the hairpin. Approximately 50% of the genomic loci were identified by alignment of both 5′ and 3′ mature miRNAs (Fig. 3b). For these genomic loci the pre-miRNA sequence lengths was estimated as the length from the 5′ miRNA start to the 3′ miRNA end. The resulting sequence lengths were plotted against the cumulative fraction of the number of pre-miRNAs, showing that the majority (>95%) of hairpins exhibited a length between 50 and 70 bases (Fig. 3c).

Since it has been shown that Drosha cleavage is dependent on the hairpin loop rather than consensus sequences in the flanking regions, the precise start and stop sites of a pre-miRNA are difficult to determine (Zeng et al., 2005). Therefore, an arbitrary distance of 10 bases upstream the 5′ miRNA and 10 bases downstream the 3′ miRNA was included as "buffer" during sequence extraction from the genomic references (Fig. 3a). Based on the observation that most pre-miRNA sequences ranged between 50 and 70 bases, an

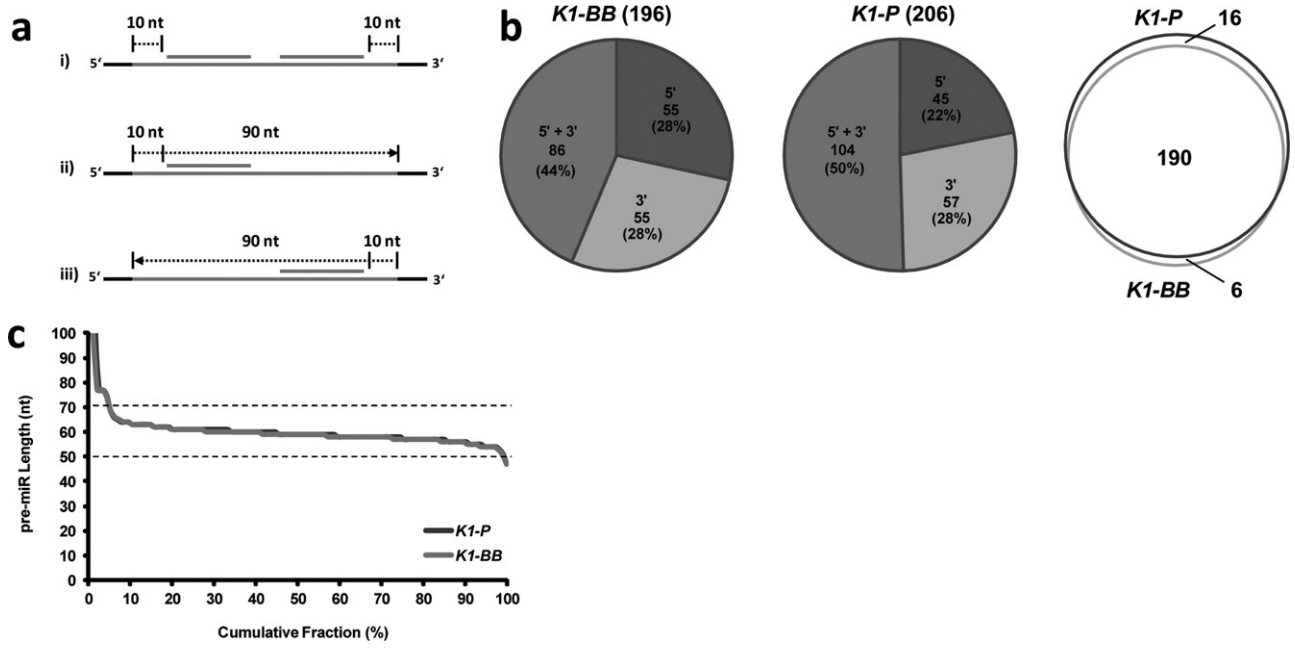

Fig. 3. Characterization of CHO pre-miRNA sequences. (a) Scheme representing the strategy for pre-miRNA sequence extraction from a genomic locus mapped by either one or two mature miRNAs: (i) a buffer of 10 bases up- and downstream the mature miRNAs was taken in case both hairpin-arms were mapped. (ii) and (iii) For genomic positions aligned by a single miRNA a total pre-miRNA of 100 bases was extracted, starting 10 bases upstream a 5′ miRNA match or 10 bases downstream a 3′ match. (b) Distribution of CHO miRNA loci identified by alignment of either 5′ or 3′ mature miRNAs or both is shown. Venn overlap of miRNA genomic loci as identified independently in each CHO-K1 genomic reference sequence. (c) For pre-miRNA genomic loci mapped at both the 5′ and 3′ miRNA hairpin-arm, length of the pre-miRNA was calculated as the distance between the start of the 5′ miRNA alignment and the end of the 3′ miRNA alignment. Cumulative fraction of pre-miRNAs is plotted against the pre-miRNA length, showing that for most pre-miRs length ranged between 50 and 70 bases.

arbitrary sequence length of 100 bases was defined for pre-miRNAs with only one expressed miRNA detected (*i.e.* only one hairpin-arm mapped by a mature miRNA), including a buffer of 10 bases upstream or downstream the miRNA start site (Fig. 3a). The important information whether a single match represented a 5′ or 3′ miRNA was derived from orthologous pre-miRNAs (mainly human, mouse or rat) in miRBase. In order to verify sequence correctness, all CHO pre-miRNA sequences were folded *in silico* using the DINAMelt webserver that is based on the mfold++ software (Markham and Zuker, 2005). Manual curation of all folding resulted in the removal of 7 putative novel CHO pre-miRs that did not resemble structures of canonical miRNAs with a complementary stemloop and 3′ overhangs, while all of the conserved CHO pre-miRs (209 sequences) as well as three novel pre-miRs passed manual curation. The respective 212 RNA secondary structures are provided as Supplementary Data 2. Table 3 exemplarily gives the pre-miRNA sequences of all 6 miRNAs belonging to the miR-17-92 cluster, which were identified in close proximity on one genomic scaffold.

Comparison of pre-miRNA sequences derived from two CHO-K1 genomic references (*K1-P* and *K1-BB*) gave four sequences with either one or two mismatches, of which only mir-486 harbored a potential single-nucleotide polymorphism (SNP) within a mature miRNA (Supplementary Table 2). In the other cases, SNPs were identified in the hairpin-loop (mir-324 and mir-486) or regions flanking mature miRNAs to the 5′ (mir-1956) or 3′ end (mir-542). Conservation of CHO pre-miRNA sequences was estimated for mir-17-92 by calculating ClustalW alignments (Thompson et al., 1994) to the respective sequences from *Mus musculus*; the results indicate high conservation for mir-18a and mir-19b, while several mismatches were found between mouse and CHO hairpin-loops of mir-17, mir-20a, and mir-92a, respectively (Fig. 4a). Supplementary Data 3 gives the sequences of all 212 miRNA hairpins, as they were extracted from the *K1-P* and *K1-BB* genomic reference as well as the respective genomic location. To show that the here provided information can easily be applied to amplify and clone CHO miRNAs, four distinct clusters of miRNAs were chosen for PCR amplification using primers designed based on the *K1-P* genomic reference (Supplementary Table 2). Genomic DNA isolated

Table 2
Number of aligned miRNAs, unique genomic loci and precursor-miRNA sequences.

	K1-P	K1-BB
miRNAs mapped to genome (100% ID, 100% length)	353	330
miRNAs mapped to genomic repeat regions	14	15
miRNAs used for identification of genomic loci and pre-miRNAs	339	315
High confidence genomic miRNA loci[a]	206	196
pre-miRNA sequences submitted to miRBase[b]	206	6

[a] After removal of loci that give rise to incorrectly folded pre-miRs.
[b] In total 212 pre-miRNA sequences submitted to miRBase.

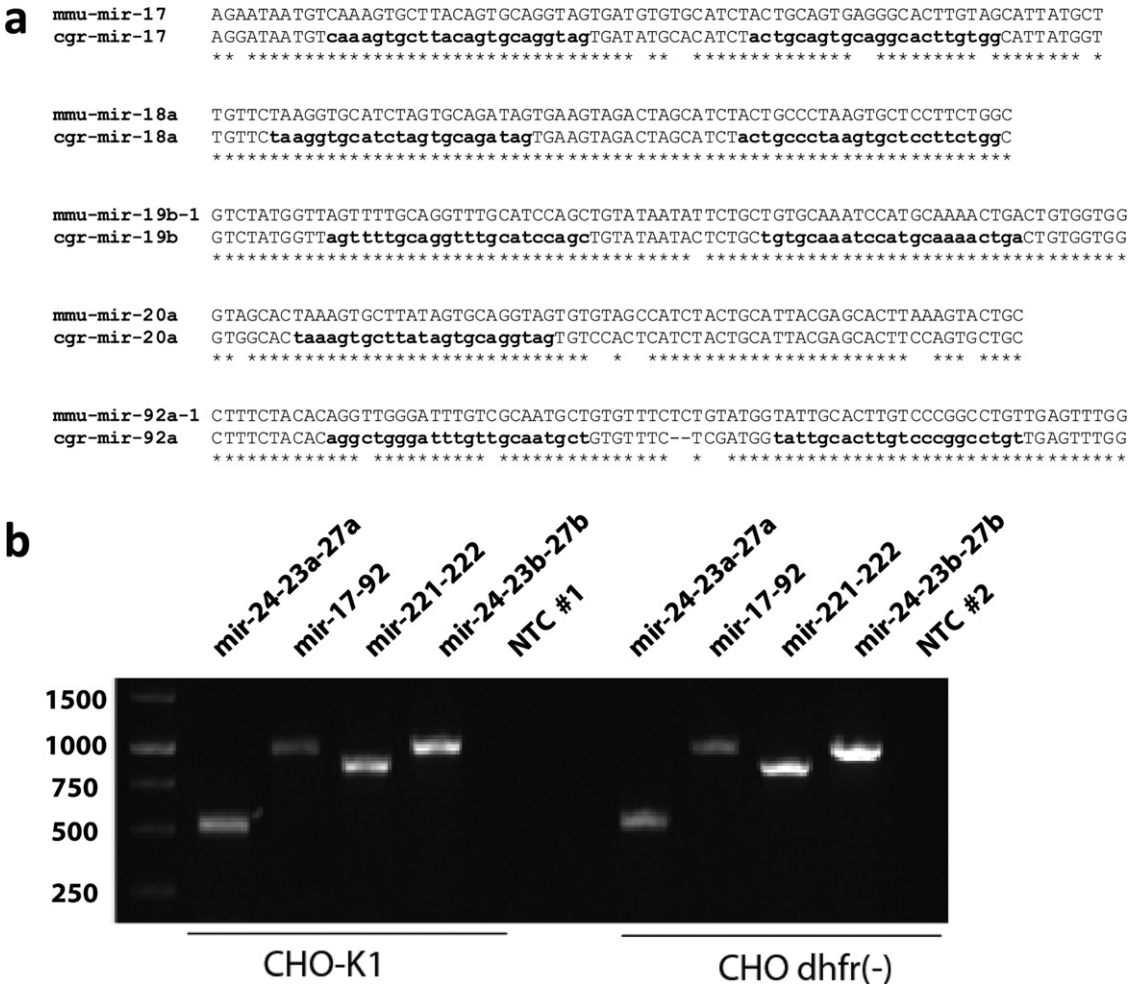

Fig. 4. Sequence characterization of CHO pre-miRNAs. (a) Conservation CHO (cgr) mir-17-92 pre-miRNAs in respect to *Mus musculus (mmu)*; *, sequence matches; - sequence deletions. (b) PCR amplification of miRNA clusters: PCR amplification of miRNA clusters using genomic DNA from CHO-K1 and CHO dhfr (-) cells. Lanes 1–4 and 6–9 showing specific amplification for miR-24-23a (1/6), miR-17-92a, miR-221-222 and miR-24-23b. Lanes 5 and 10 no template control PCR.

Table 3
miR-17-92 pre-miRNA sequences.

>cgr-mir-17_scaffold_gi|344163086|gb|JH001979.1|_REV
AGGATAATGTcaaagtgcttacagtgcaggtagTGATATGCACATCTactgcagtgcaggcacttgtggCATTATGGT
>cgr-mir-18a_scaffold_gi|344163086|gb|JH001979.1|_REV
CTTTTTGTTCtaaggtgcatctagtgcagatagTGAAGTAGACTAGCATCTactgccctaagtgctccttctggCATAAGAAG
>cgr-mir-19a_scaffold_gi|344163086|gb|JH001979.1|_REV
GCAGCCCTCTGTTAGTTTTGCATACTTGCACTACAAGAAGAATGCAGTtgtgcaaatctatgcaaaactgaTGGTGGCCT
>cgr-mir-19b_scaffold_gi|344163086|gb|JH001979.1|_REV
GTCTATGGTTagttttgcaggtttgcatccagcTGTATAATACTCTGCtgtgcaaatccatgcaaaactgaCTGTGGTGG
>cgr-mir-20a_scaffold_gi|344163086|gb|JH001979.1|_REV
TCTGTGGCACtaaagtgcttatagtgcaggtagTGTCCACTCATCTACTGCATTACGAGCACTTCCAGTGCTGCCAGCTGGAGAGCCCCAGCCTCGCTCG
>cgr-mir-92a_scaffold_gi|344163086|gb|JH001979.1|_REV
CTTTCTACACaggctgggatttgttgcaatgctGTGTTTCTCGATGGtattgcacttgtcccggcctgtTGAGTTTGG

Lower case letters indicate mature miRNAs; upper case letters indicate 5′ and 3′ flanking regions as well as loop regions.

from adherent CHO-K1 and DHFR (-) cell lines cultivated at 37 °C at 7% atmospheric CO_2, served as template for the PCR reaction that gave specific bands at the expected size (Fig. 4b).

Overall, these data demonstrate a successful identification of the genomic location of 365 out of 415 (88%) expressed mature miRNA sequences. After exclusion of 15 miRNAs due to multiple alignments to genomic repeat regions, 350 miRNAs remained for annotation of genomic loci based on miRNA alignment patterns. After manual verification of miRNA-like RNA secondary structures, a total of 212 miRNA loci as well as the respective pre-miRNA sequences were identified with high confidence (Table 2), cross-checked to confirm correctness of sequences, and provided as Supplementary Data to this publication. In addition all sequences were submitted to the miRBase database (Griffiths-Jones, 2010) for assignment of miRBase accession numbers (Supplementary Data 3).

This data can now be used to establish CHO specific tools for miRNA overexpression as engineering strategy using endogenous pre-miRNA sequences, which do show differences in nucleotide sequence compared to mouse homologs (Fig. 4b). In addition, the development of knockout strategies to specifically reduce miRNA expression will benefit from these data, and finally, knowledge of the genomic loci also allows amplification and cloning of polycistronic miRNA clusters that are likely to have a stronger influence on CHO cell phenotypes upon overexpression compared to single miRNAs.

Acknowledgments

M.H. would like to acknowledge support by the BOKU DOC scholarship, V.J., J.G. and N.B. are supported by the FWF Doctoral Program BioTop W1224, and J.G. by the GEN-AU Grant "Non-coding RNAs" 820982. T.J. acknowledges the receipt of a scholarship from the CLIB Graduate Cluster Industrial Biotechnology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbiotec.2012.01.019.

References

Barron, N., Sanchez, N., Kelly, P., Clynes, M., 2011. MicroRNAs: tiny targets for engineering CHO cell phenotypes? Biotechnol. Lett. 33, 11–21.

Becker, J., Hackl, M., Rupp, O., Jakobi, T., Schneider, J., Szczepanowski, R., Bekel, T., Borth, N., Goesmann, A., Grillari, J., Kaltschmidt, C., Noll, T., Pühler, A., Tauch, A., Brinkrolf, K., 2011. Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. J Biotechnol. 156, 227–235.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A., 2009. Rfam: updates to the RNA families database. Nucleic Acids Res. 37, D136–D140.

Gregory, R.I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., Shiekhattar, R., 2004. The microprocessor complex mediates the genesis of microRNAs. Nature 432, 235–240.

Griffiths-Jones, S., 2010. miRBase: microRNA sequences and annotation. Curr. Protoc. Bioinformatics, 1–10 (Chapter 12, Unit 12.9).

Hackl, M., Jakobi, T., Blom, J., Doppmeier, D., Brinkrolf, K., Szczepanowski, R., Bernhart, S.H., Siederdissen, C.H.Z., Bort, J.A.H., Wieser, M., Kunert, R., Jeffs, S., Hofacker, I.L., Goesmann, A., Pühler, A., Borth, N., Grillari, J., 2011. Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: identification, annotation and profiling of microRNAs as targets for cellular engineering. J. Biotechnol. 153, 62–75.

Hulsen, T., de Vlieg, J., Alkema, W., 2008. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9, 488.

Hüttenhofer, A., Schattner, P., 2006. The principles of guiding by RNA: chimeric RNA–protein enzymes. Nat. Rev. Genet. 7, 475–482.

Johnson, K.C., Jacob, N.M., Nissom, P.M., Hackl, M., Lee, L.H., Yap, M., Hu, W.-S., 2011. Conserved microRNAs in Chinese hamster ovary cell lines. Biotechnol. Bioeng. 108, 475–480.

Markham, N.R., Zuker, M., 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 33, W577–W581.

Michlewski, G., Guil, S., Semple, C.A., Cáceres, J.F., 2008. Posttranscriptional regulation of miRNAs harboring conserved terminal loops. Mol. Cell 32, 383–393.

Mudhar, P., 2006. Biopharmaceuticals: insight into today's market and a look to the future. Pharm. Technol. Europe 18, 20–25.

Müller, D., Katinger, H., Grillari, J., 2008. MicroRNAs as targets for engineering of CHO cell factories. Trends Biotechnol. 26, 359–365.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., Rajewsky, N., 2008. Widespread changes in protein synthesis induced by microRNAs. Nature 455, 58–63.

Takeshita, D., Zenno, S., Lee, W.C., Nagata, K., Saigo, K., Tanokura, M., 2007. Homodimeric structure and double-stranded RNA cleavage activity of the C-terminal RNase III domain of human dicer. J. Mol. Biol. 374, 106–120.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Xu, X., Nagarajan, H., Lewis, N.E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S., Andersen, M.R., Neff, N., Passarelli, B., Koh, W., Fan, H.C., Wang, J., Gui, Y., Lee, K.H., Betenbaugh, M.J., Quake, S.R., Famili, I., Palsson, B.O., Wang, J., 2011. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat. Biotechnol. 29, 735–741.

Zeng, Y., Yi, R., Cullen, B.R., 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. EMBO J. 24, 138–148.